Artificial intelligence in human resources management: Challenges and a path forward

Peter Cappelli,[i] Prasanna Tambe,[1] and Valery Yakubovich[ii]

Abstract

We consider the gap between the promise and reality of artificial intelligence in human resource management and suggest how progress might be made. We identify four challenges in using data science techniques for HR tasks: 1) complexity of HR phenomena, 2) constraints imposed by small data sets, 3) accountability questions associated with fairness and other ethical and legal constraints, and 4) possible adverse employee reactions to management decisions via data-based algorithms. We propose practical responses to these challenges and converge on three overlapping principles - causal reasoning, randomization and experiments, and employee contribution—that could be both economically efficient and socially appropriate for using data science in the management of employees.

Introduction

The speed with which the business rhetoric in management moved from big data (BD) to machine learning (ML) to artificial intelligence (AI) is staggering. The match between the rhetoric and reality is a different matter, however. Most companies are struggling to make any progress building data analytics capabilities: 41% percent of CEOs report that they are not at all prepared to make use of new data analytic tools, and only 4 percent say that they are "to a large extent" prepared[i].

"AI" conventionally refers to a broad class of technologies that allow a computer to perform tasks that normally require human cognition, including adaptive decision-making. Our discussion here is narrower, focusing on a sub-class of algorithms within AI in use now that rely principally on the increased availability of data for prediction tasks. For certain, there have been major advances in the domains of pattern recognition and natural language processing (NLP) over the last several years. Deep learning using neural networks has become increasingly common in some data-rich contexts and has brought us closer to true AI. Nevertheless, with respect to the management of employees, where

[i] The Wharton School, University of Pennsylvania
[ii] ESSEC Business School, France and the Center for Human Resources, the Wharton School, USA

Electronic copy available at: https://ssrn.com/abstract=3263878

the promise of more sophisticated decisions has been articulated loudly and often, few organizations have even entered the big data stage. Only 22 percent of firms say they have adopted analytics in human resources,[ii] and how sophisticated the analytics are in those firms is not at all clear.

The promise of data analytics, by contrast, is easier to see in other fields, such as operations, where the relevant questions have more clarity: for example, when will the bearings on an airplane part be likely to fail? The outcome (part failure) is clear, possibly already measured, and the number of observations may be very large (e.g., where the parts are on thousands of planes and monitored regularly) making the application of big data techniques feasible. Moreover, the idea of identifying parts before they fail seems to be unambiguously a good thing.

The effective application of AI to human resources problems presents different challenges. They range from practical to conceptual, including the fact that the nature of data science analyses when applied to people can create serious conflicts with criteria societies typically see as important for making consequential decisions about individuals. Consider the following:

1) A first problem is the complexity of HR outcomes, such as what constitutes being a "good employee." There are many dimensions to that construct, and measuring it with precision for most jobs is quite difficult: performance appraisal scores, the most widely-used metric, have been roundly criticized for problems of validity and reliability as well as for bias,[iii] and many employers are giving them up altogether.[iv] Any reasonably complex job is interdependent with other jobs and therefore individual performance is hard to disentangle from group performance. A vast literature documents numerous problems with existing performance systems as well as our field's failure to establish a clear link between individual, team, and organizational performance.[v] Given the uncertain quality of performance evaluations by humans, can we use them for training AI algorithms? Even if a high level of accuracy was achievable, this would mean scaling up arbitrary or outright discriminatory human decisions.
2) The number of times a task such as dismissals has been executed in most organizations may be in the hundreds, and while that may be sufficient for

Electronic copy available at: https://ssrn.com/abstract=3263878

managers to learn from, that number is quite small by the standards needed for data science.[vi] Moreover, many outcomes of interest are rarely observed, such as the firing of employees for poor performance, and data science techniques perform poorly when predicting relatively rare outcomes.[vii]

3) The outcomes of human resource decisions, such as who gets hired and fired, have such serious consequences for individuals and society that concerns about fairness – both procedural and distributive justice – and ethics are paramount. Elaborate legal frameworks also hold employers accountable for making those decisions in a fair manner. Central to those frameworks is the concern with explainability, knowing what attributes are driving the decision, something that is typically absent from pattern recognition methods underlying many state-of-the-art algorithms.
4) Employment decisions are also subject to a range of complex socio-psychological concerns that exist among employees, such as personal worth and status, perceived fairness, and contractual and relational expectations, that affect organizational outcomes as well as individual ones. As a result, being able to explain, justify, and get employees to accept the algorithms being used is crucial. When lacking acceptance, employees are capable of gaming or other adversarial reactions to algorithmic-based decisions that, in turn, affect organizational outcomes. While a human decision-maker can monitor adversarial behavior and adjust his or her decisions accordingly, even state-of-the-art algorithms find this to be a challenging problem. Dealing with manipulation of this type is the focus of a machine learning technique known as "adversarial machine learning".

To illustrate these concerns, consider the use of an algorithm to predict who to hire. As is typical in problems like these, the application of machine learning techniques would create an algorithm based on the attributes of employees as it relates to their job performance in the current workforce. Even if we found a causal relationship between an attribute such as sex and job performance, we might well not trust an algorithm that says hire more white men because job performance itself may be a biased indicator, the attributes of the current workforce and of our data may be distorted by how we hired in

Electronic copy available at: https://ssrn.com/abstract=3263878

the past (e.g., we hired few women), and both the legal system and social norms would create substantial problems for us if we did act on it.

In 2018, Amazon discovered that its algorithm for hiring had exactly this problem for exactly this reason. It had been built on historical job performance data, when white men had been the best performers (indeed white men were most of the employees), and the algorithm gave higher scores to white male applicants as a result. Even when the sex of applicants was not used as a criterion, attributes associated with women candidates, such as courses in "Women's Studies", caused them to be ruled out. The company soon stopped using the system as there was no simple way to fix it.[viii]

Even when we build an algorithm on a more objective measure, such as who steals from the company, the number of such cases in a typical company is too small to construct an effective algorithm. Moreover, with a task such as hiring, once applicants discover the content of our hiring algorithm, they are likely to adjust their behavior to it and render the algorithm worthless: Most applicants already know, for example, to answer the question "what is your worst characteristic" with an attribute that is not judged as negative, such as, "I work too hard."

We address below each of these challenges separately at each stage of what we call the AI Life Cycle: Operations – Data Generation – Machine Learning – Decision-Making. We recognize that researchers who study the practice of human resources have identified and discussed at length the human resource challenges faced by organizations, and we do not claim to have uncovered these challenges for the first time. The arguments below show how AI algorithms can respond to them using the approaches of contemporary data science as an alternative to managerial judgment. We introduce complex ideas from computer science and statistics into the HR context and present them in a language of managers and HR practitioners. We do not attempt to speculate as to how the practices of human resources may be changed by these data science techniques as doing so requires both a thorough review of the state of current practice and a systematic review of the barely emerging phenomenon of AI in HR management.

Substantively, we bring together key ideas from Evidence-Based Management (EBMgmt) - a theory-driven analysis of "small data"[ix] and out-of-the-mainstream approaches to machine learning[x] in order to position causation as central to all four

Electronic copy available at: https://ssrn.com/abstract=3263878

challenges we identified. We also suggest that randomization can be a useful component of an AI-augmented decision process, given that it is already present in many managers' decisions,[xi] it is often perceived as fair,[xii] and algorithms may otherwise struggle to make fair and valid decisions.[xiii]

To bridge the state-of-the-art in data science with the needs of HR practice, we brought data science faculty together with the heads of the workforce analytics function from 20 major US corporations, known for their sophisticated management systems, for a one-day workshop in the Fall of 2018. Prior to the workshop, we circulated a short survey with open-ended questions about their corporations' ongoing initiatives regarding analytics and algorithmic decision-making, barriers they face, and breakthroughs they expect. The workshop itself consisted of four sessions on the topics of data management, social media as a source of HR data, fairness and ethics of HR decisions, and employee recommendations. Each session included a presentation by a data scientist followed by an open discussion.

Our practitioners' examples and comments from the survey and workshop are not representative of the business at large. Nevertheless, they were helpful for informing our thinking and for articulating the challenges stated above.

The AI Life Cycle

Figure 1 depicts a conventional AI Life Cycle: Operations, Data Generation, Machine Learning, and Decision-Making.

FIGURE 1 HERE

"**Operations**" constitute our phenomenon of interest, such as how an organization hires employees. One of the reasons for the interest in applying data science tools to human resource operations is because HR performs so many tasks and so much money is involved in them. In the US economy as a whole, roughly 60 percent of all spending is on labor. In service industries, the figure is much higher.[xiv] In the table below, we list the most common operations in human resources with corresponding prediction tasks for workforce analytics. They correspond to the "Human Resources Life Cycle," which is commonly used to organize HR tasks.[xv] :

HR operation	**Prediction task**

Electronic copy available at: https://ssrn.com/abstract=3263878

Recruiting – identifying possible candidates and persuading them to apply	Are we securing good candidates?
Selection – choosing which candidate should receive job offers	Are we offering jobs to those who will be the best employees?
On-boarding - bringing an employee into an organization	Which practices cause new hires to become useful faster?
Training	What interventions make sense for which individuals, and do they improve performance?
Performance management – identifying good and bad performance	Do our practices improve job performance?
Advancement – determining who gets promoted	Can we predict who will perform best in new roles?
Retention	Can we predict who is likely to leave and manage the level of retention?
Employee benefits	Can we identify which benefits matter most to employees to know what to give them and what to recommend when there are choices, and what are the effects of those benefits (e.g., do they improve recruiting and retention)?

Each of these operations involves administrative tasks, each affects the performance of the organization in important ways, and each includes specific offices, job roles, written instructions and guidelines to execute as well as the actual activities and interactions of all parties. These operations produce volumes of data, in the form of texts, recordings, and other artifacts. As operations move to the virtual space, many of these outputs are in the form of "digital exhaust," which is trace data on digital activities (e.g. online job applications, skills assessment) that may be used to build recruiting algorithms.

Human resource information systems, applicant tracking systems, digital exhaust, and other markers are all critical inputs for the "**data generation"** stage. Typically, this

Electronic copy available at: https://ssrn.com/abstract=3263878

input has to be extracted from multiple databases, converted to a common format, and joined together before analysis can take place.

By "**machine learning,**" (ML) we refer to a broad set of techniques that can adapt and learn from data to create algorithms that perform better and better at a task, typically prediction. Within business contexts, the most common application of machine learning technologies has been "supervised" applications, in which a data scientist trains a machine learning algorithm on a labeled training sample and determines the most appropriate metric to assess its accuracy. Some of the most commonly used prediction algorithms, such as "logistic regression", infer the outcome variable of interest from statistical correlations among observed variables[xvi].

For hiring, for example, we might see which applicant characteristics have been associated with better job performance and use that to select candidates in the future. "Algorithmic management," the practice of using algorithms to guide incentives and other tools for "nudging" platform workers and contractors in the direction of the contractee[xvii] is applied to regular employees.[xviii] At present, this is principally the case in making recommendations to employees about actions they may take. IBM, for example, uses algorithms to advise employees on what training make sense for them to take, based on the experiences of similar employees; the vendor Quine uses the career progression of previous employees to make recommendations to client's employees about which career moves make sense for them.

Vendors such as Benefitfocus develop customized recommendations for employee benefits, much in the same way that Netflix recommends content based on consumer preferences or Amazon recommends products based on purchasing or browsing behavior. The extension of such recommendations into wellness programs is already underway, in some cases collecting data about employees' health and wellness directly with devices like "Fitbits," urging employees to adopt practices that lead to better health outcomes, and sometimes rewarding and punishing them with payments or higher healthcare costs based on their compliance.

These algorithms differ in some important ways from traditional approaches used in HR. In industrial psychology, the field that historically focused the most attention on human resource decisions, research on hiring, say, would test separate explanatory

Electronic copy available at: https://ssrn.com/abstract=3263878

hypotheses about the relationship between individual predictors and job performance. The researcher picks the hypothesis to examine and the variables with which to examine it. This process produces lessons for hiring, typically one test at a time, e.g., the relationship between personality test scores and job performance, then the relationship between education and job performance, and so forth. The result would be conclusions about several variables that might be used to predict hiring success.

Machine learning, in contrast, uses many variables to generate one algorithm and typically one score to assess a candidate. The variables used may not be in the cannon of the theoretical literature associated with the topic, and the researcher is not hypothesizing or indeed even examining the relationship between any one variable and the outcome being predicted. Indeed, one of the attractions of ML is its investigation of non-traditional factors because the goal is to build a better prediction rather than advancing the theory of the field in which the researcher is based by providing evidence on particular hypotheses.

The second most popular use of data science in human resources may be in predicting turnover. Vendors like Jobvite generate machine learning algorithms that score individual employees based on social media posts; others use simpler data like the extent to which individuals have updated their LinkedIn profiles. Many of the companies at our conference were developing their own, proprietary algorithms to predict flight risk.

IBM's Blue Match software uses algorithms in a more novel manner, to drive career advancement by suggesting career advancement moves and new jobs to apply for in the company based on employee interests and prior jobs, training, and ultimately the characteristics of individuals who have succeeded in those jobs in the past. Twenty-seven percent of the company's employees who changed jobs in 2018 did so based on recommendations from the company's Blue Match software.[xix]

The move away from check-list based performance appraisals and toward continuous discussions, facilitated by phone-based apps, has been facilitated by natural language processing software from vendors like Work Compass. These systems read through a year's worth of text messages to produce summaries of the issues discussed and comparisons with other employees, among other things, to drive merit pay decisions.

"**Decision-making**," the final stage, deals with the way in which we use insights from the machine learning model in everyday operations. In the area of human resource

Electronic copy available at: https://ssrn.com/abstract=3263878

decisions, individual managers may have more discretion now in how they use empirical evidence from data science and other models than they did in the heyday of the great corporations when hiring and other practices were standardized across an entire company. Managers today often have the option of ignoring predictors about candidate success, for example, using it as they see fit, and generating their own data in the form of interviews they structure themselves.

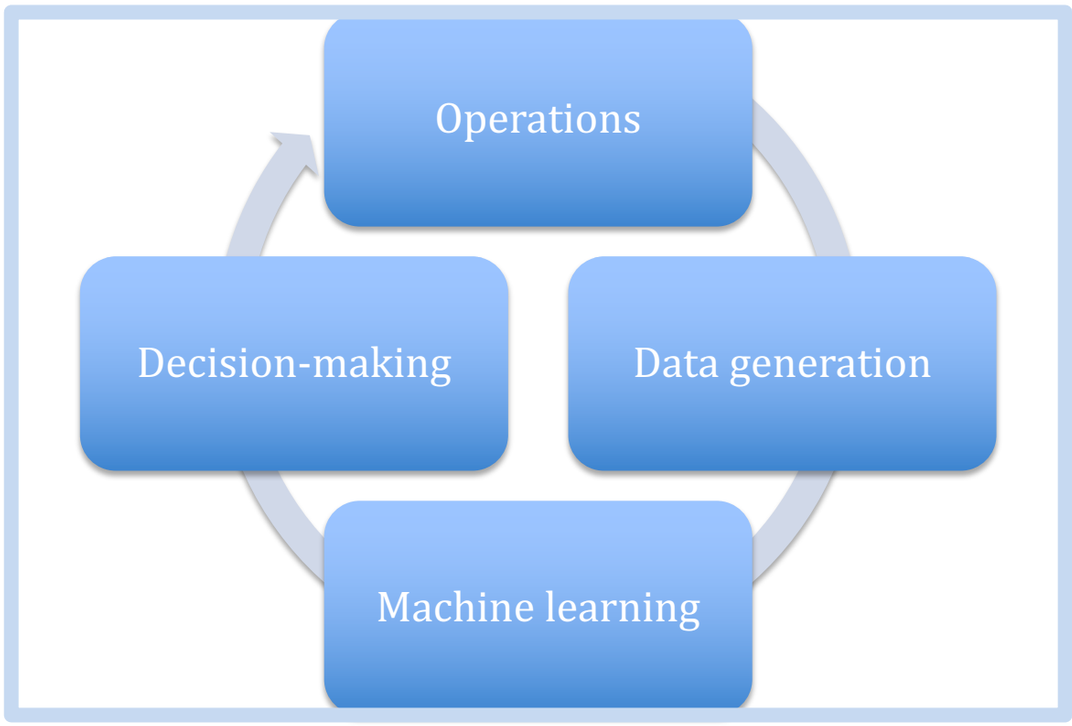

Figure 1. The life cycle of an AI-supported HR practice

Addressing AI Challenges: One Stage at a Time

In this section, we explore in detail the four general challenges to AI outlined in the Introduction: complexity of HR phenomena, small data, ethical and legal constraints, and employee reactions to AI-management. To make these challenges tractable, we discuss them in the context of the particular stages of the AI Life Cycle in which they are most relevant.

Data Generation stage

The *complexity* inherent in many HR phenomena manifests itself at the Data Generation stage. The most important source of complexity may be the fact that it is not easy to measure what constitutes a "good employee," given that job requirements are

Electronic copy available at: https://ssrn.com/abstract=3263878

broad, monitoring of work outcomes is poor, and biases associated with assessing individual performance are legion.[xx] Moreover, complex jobs are interdependent with one another, and thus one employee's performance is often inextricable from the performance of the group: Is it sufficient to be a good individual contributor, and if not, how do we measure interactions with others? Without clear measures of what it means to be a good employee, a great many HR operations face considerable difficulty in measuring performance, which is the outcome driving many HR decisions.

In terms of the data, the participants of our workshop indicated that not all attributes of HR actions that we imagine are measured actually are; not all details of operations leave digital traces that could be captured, and not all traces left can be extracted and converted to a usable format at a reasonable cost. For example, employers do not necessarily track the channels through which applicants come to them – from referrals vs. visiting our website vs. job boards, and so forth – which is a reasonably simple exercise to do. Most employers collect only a limited amount of data on applicants before ruling them out, and they do not retain it for those applicants that they screen out. These choices limit the types of analyses that can be performed and the conclusions that can be drawn.

The fact that there is no list of "standard" variables that employers are required to gather and retain through their HR operations, as there might be in fields like accounting, reduces the extent to which best practices in analytics can be transferred across organizations. Behavioral measures from attitudinal surveys, for example, vary considerably across organizations, measures of job performance differ, differences in cost accounting mean that the detail that employers have on the costs of employees differs enormously (e.g., are training costs tracked, and if so, are they aggregated in ways that limit the ability to examine them?), and so forth.

When tackling the challenge of data generation, employers can benefit from the lessons drawn from fields like performance management:

- Do not expect perfect measures of performance as they do not exist. It is better to choose reasonable measures (e.g., would you have hired this new employee if you could go back) and stick with them to see patterns and changes in results than to keep tinkering with systems to find the perfect measure. Most of our data

Electronic copy available at: https://ssrn.com/abstract=3263878

analytics efforts in HR are based on decisions concerning individual employees – who to hire, who to retain, what to recommend about training and advancement - we need good measures of what constitutes success, who is a good performer, or none of these efforts will succeed.

- Objective measures of performance outcomes based on ex-ante determined goals and Key Performance Indicators are best, but they are never complete. Complement them with measures to capture less tangible outcomes, such as whether the employee fits into the company's culture, even if those measures are subjective, to prevent a situation where employees optimize on the few objective measures at the expense of everything else.
- Include business and financial performance data at the organizational level closest to employee control to have the best chance of seeing how individual performance affects larger business units and the company as a whole.
- Aggregate information from multiple perspectives and over time. Digital HR tools allow for rapid real-time evaluations among colleagues using mobile devices, for example. Machine learning algorithms are ideal for making sense of such information.

The complexity of HR phenomena creates another problem in the form of specialized vendors who address only one task. It is very common for an employer to have a system from one vendor to track employee performance scores, from another for applicant tracking software, from a third for compensation and payroll data, and so forth. The biggest practical challenge in the data generation phase and arguably in using data in human resources at all is simply database management, aggregating existing data so that it can be examined because the systems are rarely compatible. It is no surprise that such database challenges were one of the biggest challenges reported by the HR analytics practitioners in our workshop. In addition to technical barriers, our respondents reported the resistance of other functions to sharing their data with HR Departments.

To illustrate how rudimentary most of the existing database management efforts still are with HR operations, the vast majority of our practitioners reported that the software they most often used to organize, manage, and analyze their data was Excel. Very few used more purpose-built tools such as Tableau that are common in data

Electronic copy available at: https://ssrn.com/abstract=3263878

analytics. Software for bridging datasets and "data lakes" that can archive and access different data sets represent a way forward, but they can be difficult to integrate, can be viewed as confining, and face their own limitations, so they remain under-used in the HR world.

To demonstrate its commitment to digital transformation as well as to benefit from it, companies' top management has to make data sharing a priority in the short-run and invest in data standardization and platform integration in the long-run. The fact that at present, the separate types of data needed to do even the most basic analyses, such as seeing whether hiring decisions lead to better new employees, cannot be done because the components of data are owned by different parts of the organization. Only executives at a higher level can make the cooperation across units happen that is a necessary condition before data analysis can begin.

Given these database concerns, it can be extremely difficult and costly to analyze a question in HR for the first time. Data analytics managers, therefore, have to be careful about where to "place bets" in terms of assembling data for analysis, let alone when collecting new data. How should managers decide which HR questions to investigate, especially when so few have been examined before?

This challenge was the most important concern in our discussion with practitioners. Beyond the obvious criteria of cost is the likelihood of generating useable results. Our practitioners said that in this context, they relied on induction to make the choice: they ask people in HR operations what they have seen and what they think the important relationships are. Some go to senior management and solicit answers to the question of what types of problems prevent the managers from "sleeping at night." Such experience-driven heuristics are a typical approach under uncertainty. The practitioners also indicated that another factor shaping where they placed their bets is whether anyone was willing to act on results they found.

A more systematic response would include examining the research literature in order to establish what is already known about different research questions, as evidence-based management has long advocated.[xxi] The fact that this approach appears not to be used very often reflects the disconnect between the data science community, which understands analytics but not HR, and the HR community, which understands HR but not

Electronic copy available at: https://ssrn.com/abstract=3263878

analytics. Many leading IT companies, such as Amazon, Google, Facebook, and Microsoft, hire as many PhDs in social sciences as in data sciences into the HR department to help close this disconnect.

The last step in the process of deciding what to analyze is with an audit of what data are necessary to answer the research question and how difficult it is to assemble. For example, if the employer wants to use a machine-learning algorithm in hiring, it needs to have historical data on job candidates who were not hired, something that many employers do not retain. It may not be possible to answer questions that are important and that data science is well-suited to answer because the data are not available.

Small data is a fundamental concern for human resource analytics. Most employers do not hire many workers, nor do they do enough performance appraisals or collect enough other data points for their current workforce to use machine learning techniques because they do not have that many employees. The machine learning literature has shown that access to larger data has substantial advantages in terms of predictive accuracy.

At the same time, even if data sets are not big enough for machine learning exercises, small data are often sufficient for identifying relationships; we may not be able to build a machine learning algorithm for hiring, but we probably do have enough data to answer questions about specific hiring criteria, such as whether recruiting from the CEO's Alma Mater really produces better hires. Some aspects of HR may generate millions of observations, such as continuous measures of employee performance. It is straight-forward, for example, to monitor employee time spent doing online work and not working; call center employees are assessed on each call with many metrics; employees performing simple physical tasks, such as sorting packages, are measured per hand movement.[xxii]

The management literature has an important advantage over data science in articulating causal relationships, as opposed to prediction from correlations among observed variables in machine learning. Recently, voices in the computer science community have articulated the problem of causation as critical for the future of AI in human affairs.[xxiii] We consider the issue of causality in more detail below.

Electronic copy available at: https://ssrn.com/abstract=3263878

The less data we have, the less we can learn from data analytics, and the more we need from theory and prior research to identify causal predictors of the outcome of interest. AI-management requires that managers put their assumptions on the table, though, and persuade the other stakeholders about their accuracy, ultimately by using data and empirical analysis. The formulation of such assumptions often turns into a contest among stakeholders. This is a place where process formalization that presumes contributions from stakeholders is required.

Where a formal process reveals large disagreements as to causal factors, a way forward might include generating additional data from randomized experiments in order to test causal assumptions. Google became known for running experiments for all kinds of HR phenomena, from the optimal number of interviews per job candidate to the optimal size of the dinner plate in the cafeteria.[xxiv] (Off-the-record conversations also suggest that Google leadership did not accept the research finding that unstructured interviews were poor predictors of good hires – having committed to that practice - and so conducted the research that confirmed it was true even at Google.) If discussions, experiments, and leadership's persuasion do not lead to a reasonable consensus on the causal model that generates the outcome of interest, AI-analyses are likely to be counterproductive.

One attraction of using vendors is their ability to combine data from many employers to generate their algorithms. Such approaches have long been used with standard paper-and-pencil selection tests, or as they are sometimes known now, pre-employment tests, such as those for sales roles. For instance, the company ADP, which handles outsourced payroll operations for thousands of companies, has been able to harness this scale to build predictive models of compensation and churn. Client companies are willing to make their data available for this exercise in return for access to the predictive models and benchmarked comparisons.

The complication for individual employers is knowing to what extent their context is distinct enough that an algorithm built on data from elsewhere will make effective predictions in their own organization. As is discussed further below, such evidence is essential to address legal concerns.

Electronic copy available at: https://ssrn.com/abstract=3263878

Employers are also concerned about employees' tendency to bias their responses and the data depending on how they think the data are used. Because of this, a great many employers now make use of social media information precisely because they believe employees are being more authentic in it.[xxv] That data gets used in hiring (e.g., looking for evidence of bad behavior, looking for evidence of fit) and to assess "flight risk" or retention problems (e.g., identifying updated LinkedIn profiles). Banks have tighter regulations requiring oversight of employees and have long analyzed email data for evidence of fraudulent behavior. They are now using it as well to identify other problems. For example, the appearance of terms like "harassment" in email traffic may well trigger an internal investigation to spot problems in the workplace.

The vendor Vibe, for example, uses natural language processing tools to gauge the tone of comments that employees post on internal chat boards, thereby helping to predict employee flight risk. Applications such as these can face some key challenges when introduced into the workplace. For instance, when employees realize their posts are being used to derive these types of measures, it can influence what and how they choose to post. Then, there are the issues that may arise around whether employees consider such use of the data to infringe upon their privacy.

Several of the companies at our workshop reported that they built models on predicting flight risk and that the best predictors did not come from traditional psychology-based findings but from data sources like social media. Many employers felt that there was an ethical problem with their own use of social media; others felt that data was ok to use but that tracking sentiment on email messages using natural language algorithms was out of bounds; still others thought that any employee-related data was appropriate to use as long as it was anonymized.

Many of these and similar considerations fall under the purview of privacy. Issues associated with electronic monitoring of employee performance and privacy are not new,[xxvi] but the contemporary context of social media in particular creates new challenges (Tucker 2017): Data can persist well beyond the time it was generated, employers can repurpose it for use unanticipated by the creator, e.g., the words from an email exchange with a colleague might be used to predict flight risk. Data of one person may also inadvertently affect other people, for example, the creators' friends tagged in posts and

Electronic copy available at: https://ssrn.com/abstract=3263878

photos. Here employers have to account for governments' regulations of privacy issues, such as "the right to be forgotten" or the EU's General Data Protection Regulation (GDPR). The former states that business has to satisfy individuals' demands to delete their digital traces after some period of time; the latter is a comprehensive treatment of all the aspects of data privacy in the digital age.[xxvii] Among novel suggestions are that the Genetic Information Nondiscrimination Act be used as a model for protecting employees their employer's breach of privacy.[xxviii]

In terms of technological solutions to the issue of data privacy, computer scientists are actively working on privacy-preserving data analytic methods that rely on the notion of differential privacy in building algorithms. Here, data is randomized during the collection process, which leads to "learning nothing about an individual while learning useful information about the population" [xxix]. Analysts do not know whose data are used in the analysis and whose data are replaced by noise, but they do know the noise generating procedure and thus can estimate the model anyway.

The practical problem with using "authentic" data, such as that in email traffic or on social media, is that it is not clear how "authentic" it really is. It is certainly true that individuals are not necessarily shaping their social media entries with the goal of influencing employers, but few people would believe that those entries are necessarily authentic. They are typically designed to create an image of the individual that is different from reality: entries about vacation cruises far outnumber entries about doing the laundry even though most of us spend far more time on the latter than the former.

The issue of individuals and especially job applicants altering their responses to what they believe assessments want, or faking, is not new[xxx]. In the case of social media data, the nature of what employees post will no doubt change as soon as individuals recognize that employers are monitoring those entries: expect far more entries about self-improvement, achievements at work, and so forth. Efforts to use computer games to assess candidates is yet another effort to obtain authentic data where the employees do not necessarily know how to spin their responses. But they are already getting help from businesses like the JobTestPrep company that helps potential candidates for jobs at Lloyds Bank figure out how to score well on Lloyds' selection game[xxxi]. Getting

Electronic copy available at: https://ssrn.com/abstract=3263878

authentic data on applicants will remain a challenge because of the ability of candidates to game such efforts.

Machine Learning stage

An ML algorithm for predicting which candidates to hire may well perform better than anything an employer has used before. Indeed, a reasonable complaint is that prior research in human resources is not making much progress to help employers: the fact that most of the predictors advocated in that research on a topic like hiring, such as personality, predict so little of job performance (a typical validity coefficient of .30, for example, translates to explaining nine percent of the variance in performance) that it creates an enormous opportunity for data analytics to do better. It will, because its goal is just to predict, and it is not limited to a small number of one-at-a-time results, such as a personality test, nor is it constrained by prior research findings

Bo Cowgill, for example, shows how an ML algorithm can do better than humans. In a field experiment with hiring white-collar workers, he finds that AI can remove human biases if the training data are sufficiently noisy: Inconsistent human decision-making introduces quasi-experimental variation which improves machine learning to such a degree that it yields better candidates than HR staff. Specifically, the candidates selected by the machine are 14% more likely to pass interviews and receive a job offer, 18% more likely to accept an extended job offer, are 0.2–0.4 standard deviation more productive once hired, and 12% less likely to show evidence of competing job offers during salary negotiations. Surprisingly, these improved results were due to noisy, inconsistent training data from hiring "non-traditional" candidates from non-elite colleges, lacking job-referrals and prior experience, with atypical credentials and strong non-cognitive soft-skills. These are remarkable counterintuitive findings that attest to the potential of AI.[xxxii]

As noted above, finding good data with which to build an algorithm can be challenging. Because clients rarely have data on employee performance in which they feel confident, a common approach in the vendor community is to build an algorithm based on the attributes of a client firm's "best performers," which are easier to identify. Then applicants are assessed against that algorithm. Consider, for example, vendors like

Electronic copy available at: https://ssrn.com/abstract=3263878

HireVue that help clients conduct video interviews. Part of their offerings include algorithms based on facial expressions captured on those videos. These algorithms are sometimes trained on data from top performers at the client firm, and job candidates are assessed based on how similar their expressions are to those of the algorithm.

Is it possible that facial expressions actually predict job performance? Social scientists may find examples like this absurd because there is no reason to expect such a relationship. The machine learning models and the data scientists behind them, of course, do not care whether we know what the reason might be for such a relationship or whether it corresponds with what we know from research on humans. They only care if there is such a relationship.

Examples like this algorithm raise many concerns, though, even for the basic goal of producing an effective algorithm. First, they run the danger of "selecting on the dependent variable" by examining only those who are successful. The algorithm may well capture attributes of good performers accurately, but it is not identifying whether those attributes are truly distinct from those of other performers. Good performers and bad performers may have the same expressions in response to situations, but we will never know without examining both groups.

The use of an algorithm or indeed any decision rule in hiring is a challenge for the "learning" aspect of machine learning because of the sample selection it generates: Once we rule out hiring candidates who are not chosen by the algorithm, the opportunity to see whether other attributes might lead to better performers diminishes and may end – say if job requirements change or if new attributes appear among candidates. In other words, the opportunity for the machine learning algorithm to keep learning disappears if we use only that algorithm to drive hiring decisions. The only way to avoid this problem is to on occasion turn off the algorithm, to not use it to hire, in order to see whether candidates that do not fit its criteria continue to perform worse or perhaps perform better.

This problem that selection based on the hiring criteria prevents learning about that criteria holds for any criterion. With the more standard hiring practice of using only a few selection criteria, it is possible to turn them off one-at-a-time to see the effect, for example, of recruiting from a different set of schools. An algorithm generated by

Electronic copy available at: https://ssrn.com/abstract=3263878

machine learning operates as one entity rolling many variables together into an overall model. As a result, it is much more difficult to turn off just one criterion.

Selection can also induce a type of spurious relationship among workers' characteristics called the collider effect in epidemiology and in data science.[xxxiii] It occurs when samples are selected in ways that restrict the range of the variables, sometimes known as "range restriction" in psychology. An employer who selects new hires based on college grades and conscientiousness tests might well find that candidates who have neither good grades nor good scores on conscientious tests are not hired. When the employer looks for a relationship between college grades and conscientiousness among its employees, it finds the relationship is negative, even though in the broader population the relationship is positive.

More generally, this selection process can reduce the range on variables of interest, making it more difficult to find true effects. For example, if we only hire candidates with good college grades, it may be difficult to identify a true, positive relationship between grades and job performance because the variance of grades in the sample is too limited to identify that relationship. Range restriction also happens when applicants self-select into a firm's pool of applicants, the first step in the well-known "attraction-selection-attrition" framework.[xxxiv](Schneider 1987). Algorithms that are based solely on data from the current workforce create this problem as well.

Several aspects of the modeling process per se can also be challenging. For instance, there is more than one measure of "fit" with the data. A well-known case of this problem concerned the use of a machine learning algorithm by judges in Broward County, Florida to determine whether a person charged with a crime should be released on bail. The algorithm was trained based on data about whether parolees violated the terms of their parole. The challenge in the data is that the majority of the individuals in the dataset were white, and so the algorithm was driven largely by information about whites. The algorithm predicted the rate of recidivism correctly at an equal rate for whites and blacks, but when it did not predict accurately, it was far more likely to over predict for blacks than for whites.[xxxv] The problem is that the algorithm cannot optimize on more than one measure of fit. The implications for human resources are obvious

Electronic copy available at: https://ssrn.com/abstract=3263878

given that prediction models for hiring or other outcomes may differ by sex, race, and other protected groups.

Decision-Making stage

There are three main challenges when decision makers try to apply the predictions produced by machine learning. The first concerns fairness and legal issues, the second relates to a lack of explainability of the algorithm, and the third to the question of how employees will react to algorithmic decisions.

Fairness

Within the HR context, there are numerous questions related to fairness. One of the most obvious of these is the recognition that any algorithm is likely to be backward looking. The presence of past discrimination in the data used to build a hiring algorithm, for example, is likely to lead to a model that may disproportionately select on white males. Actions using those algorithms risk reproducing the demographic diversity – or lack thereof - that exists in the historical data. The biased outcomes of the Amazon hiring algorithm noted above was caused by exactly this common problem: because fewer women were hired in the past and because men had higher performance scores, the algorithm was selecting out women and those with attributes associated with women.

In the HR context, there is a wide-spread belief that evaluations of candidates and employees are shaped heavily by the biases of the evaluator, most commonly as related to demographics. Algorithms can reduce that bias by standardizing the application of criteria to outcomes and by removing information that is irrelevant to performance but that might influence hiring manager decisions, such as the race and sex of candidates. On the other hand, factors that may seem inappropriate may nonetheless improve the predictive power of the algorithms, such as the social status of one's alma mater. How we balance the trade-off between appropriateness and predictive power is not clear.

The fact that employment decisions are so important to individual candidates/employees and to broader society has led to an extensive legal framework designed to guide those decisions. The vast majority of individuals in the US labor force – everyone other than white men under age 40 who do not have disabilities or relevant medical conditions – are protected against discrimination in any employment decision.

Electronic copy available at: https://ssrn.com/abstract=3263878

Other countries have similar rules. Discrimination means adverse actions taken based on one's demographic attributes, and in practice that is measured by "adverse impact," evidence that any employer's decisions have a lower incidence of good outcomes (e.g., hires and promotions) and/or a higher incidence of bad outcomes (e.g., dismissals) than the base rate we would expect from their distribution in the relevant population.[xxxvi]

With respect to the actions that could be based on algorithms, in other words, those that attempt to predict future outcomes, the only defense against evidence of adverse impact is first to show that the decisions taken actually do predict the desired outcomes and second to show that no other process for making decisions would produce at least as accurate predictions with less adverse impact.

These legal constraints raise considerable challenges for algorithm-based employment decisions. The first is simply that in order to assess whether they have an adverse impact, we have to identify the relationships within the algorithm between any of the attributes of protected groups and the relevant outcomes: Does it give women a lower score, for example, or does it give lower scores to attributes disproportionately associated with women? This is a considerable analytic task for most algorithms.

Letting supervisors make employment decisions without guidance, on the other hand, may well lead to far more bias and possibly more adverse impact than the algorithms generate. But that bias is much harder to hold accountable because it is unsystematic and specific to each hiring manager. Algorithms used across the entire organization may have less bias than relying on disparate supervisors, but bias that does result is easier to identify and affects entire classes of individuals. All of this makes it much easier to challenge hiring decisions based on algorithms. Will employers find it worthwhile to take on greater legal risk in order to reduce total bias? How will the courts consider evidence concerning algorithms in these decisions? So far, we have no experience on these issues.

If we return to the parole violation example above, it would seem that a better approach to building an algorithm to predict parole violations would be to generate a separate one for blacks and for whites. In the context of HR decisions, that might seem appealing as well, to generate separate hiring algorithms, for example, for men and women. While there may be challenges in using such algorithms (e.g., how do we

Electronic copy available at: https://ssrn.com/abstract=3263878

compare the scores of these two different models?), the legal frameworks will not allow us to treat these demographic groups differently.

These examples raise the more general concern about fundamental tradeoffs between accuracy and fairness that must be confronted in any HR machine learning implementation.[xxxvii] Consider how the role of context changes our judgments. Most of the participants at our workshop, for example, found it perfectly acceptable to use algorithms to make decisions that essentially reward employees – who to promote, who to hire in the first place. But what about the inevitable use of algorithms to punish employees? An algorithm that predicts future contributions will most certainly be introduced at some point to make layoff decisions. How about one that predicts who will steal from the company or commit a crime? Such "integrity" tests are already used in the workplace now as part of the hiring process. [xxxviii]

We see two approaches that can make progress on at least some of the above issues. The first and arguably most comprehensive approach is causal discovery, that is, identifying in the data those variables that truly cause the outcome of interest, such as good job performance. This is a fundamental distinction between data science as it is most often applied to generating algorithms that are valued principally for their predictive accuracy and conventional statistics.

Consider the question as to whether the social status of an applicant's alma mater predicts their job performance if they were hired. From the perspective of generating algorithms, it is enough if the social status measure contributes to the overall accuracy of an algorithm predicting job performance. Traditional statistics, on the other hand, might ask whether the relationship between social status and job performance is true on its own – not just as part of a more complex algorithm – and whether it was causal. Establishing causation is a much more difficult exercise.

Demonstrably causal algorithms are more defendable in the court of law and thus address at least some legal constraints discussed above. They are fairer due to the explicit specification of causal paths from socio-demographic characteristics to performance, which allows individuals to be acknowledged for their performance enhancing characteristics (e.g., grit or intrinsic motivation) independently of group membership (e.g., the alma mater status) and to intervene in order to compensate for their socio-

Electronic copy available at: https://ssrn.com/abstract=3263878

demographic disadvantages (e.g., to create a strong support network that graduates from top schools get by default). As a result, employees "minimize or eliminate the causal dependence on factors outside an individual's control, such as their perceived race or where they were born,"[xxxix] and thus are treated as individuals rather than group members. Individual fairness, in this case, replaces group fairness.

Computer algorithms can assist in causal discovery by searching for causal diagrams that fit the available data. Such algorithms are being actively developed; their interpretation does not require advanced training but does require data about possible causes and their confounders.[xl] As noted above, when data are incomplete, one can test for the causality of specific factors with randomized field experiments.

Instead of boosting the low predictive power of many HR algorithms with non-causal covariates, which exacerbate unfairness, we propose to accept that some HR outcomes are often random, or at least have random aspects to them. As noted above, Cowgill shows that noise and inconsistency in human decision-making regarding HR creates variation that can actually be used to de-bias algorithms.[xli] This is because, when we have valid long-term outcome measures (e.g. of longer-term employee performance metrics), noise can serve to "experimentally" select in observations in an earlier stage that the algorithm may have otherwise removed from the consideration set due to bias. If these observations perform well in terms of their later stage outcomes, this information can be fed back to the model to increase the likelihood they get selected in the earlier stage.

Research shows that employees understand the random aspect of many outcomes and perceive explicitly random decisions as fair in determining complex and thus uncertain outcomes.[xlii] "Flipping a coin" has a long history as a device for settling disputes, from ties in election outcomes to allocating fishing rights.[xliii] Introducing explicit randomization in decisions is especially attractive where there are "losers" in the outcomes and where they remain in the organization or relationship, such as employees who are not selected for promotion. Telling them that the decision literally was made on a coin toss is much easier to bear than either telling them it was a close choice (you were almost as good, on the one hand, but something small could have changed the outcome)

Electronic copy available at: https://ssrn.com/abstract=3263878

or that it was not close (you were not almost as good, but there is nothing you could have done that would have mattered).

It might also be helpful to introduce something less than complete randomness to the process to help with its acceptability. For example, when predictive scores are not tied but are merely close, we might introduce a weighted random aspect where the candidate with the higher score gets a proportionately greater chance. The common use of "cut scores" in tests where we assume that everyone who scored above a stated standard has "passed" and those below "failed" categories is one example where we might select winners at random from those who passed the standard.

Explainability

Closely related to the notion of fairness is explainability, in this case the extent to which employees understand the criteria used for data analytic-based decisions. A simple seniority decision rule – more senior workers get preference over less senior ones – is easy to understand and feels objective even if we do not always like its implications. A machine learning algorithm based on a weighted combination of 10 performance-related factors is much more difficult to understand, especially when employees make inevitable comparisons with each other and cannot see the basis of different outcomes. (Professors who have to explain to students why their grade is different than that of their friend who they believe wrote a similar answer are familiar with this problem.) Algorithms get more accurate the more complicated they are, but they also become more difficult to understand and explain.

A well-known example of the importance of explainability to users comes from the well-known application of algorithms to oncology by IBM Watson. This algorithm that was developed to identify cases of cancer met considerable resistance from oncologists because it was difficult to understand how the system was arriving at its decisions. When the application disagreed with the doctor's assessment, this lack of transparency made it difficult for medical experts to accept and act upon the recommendations that the system produced.[xliv] Especially in "high stakes" contexts, such as those that affect people's lives—or their careers--explainability is likely to become imperative for the successful use of machine learning technologies. We expect major

Electronic copy available at: https://ssrn.com/abstract=3263878

progress in this area in the coming years, due to a wave of investment from the commercial and government sectors geared towards explainable AI. For instance, the US Defense Advanced Research Projects Agency (DARPA), known for its successful funding of path-breaking research in IT, has just launched a major initiative on explainable artificial intelligence (XAI) with deliverables, software toolkits and computational models, expected by 2021.[xlv]

Back to Operations: Employee reactions to algorithmic decisions

Changes in formal decision-making of the kind associated with the introduction of algorithms unavoidably affect employees' experiences and behavior. In this regard, we can learn a great deal from Scientific Management's efforts to develop optimal workplace decision rules in the previous century. Employment practices (e.g., how fast to work based on time and motion studies) and decisions about work organization (e.g., breaking down tasks to simple components) were based on a priori engineering principles and human experiments. Although they may have been much more efficient than previous practices, they were bitterly resented by workers, leading to a generation of strife and conflict between workers and management. From the perspective of front-line workers and their supervisors, the situation may have looked very similar to the AI model we outline here: decisions would been handed down from another department in the organization, the justification for them would be that they were the most efficient that science could provide, understanding the basis of the decision is extremely difficult, and trying to alter them would simply be a mistake.

To illustrate, it is widely believed that the relationship with one's supervisor is crucial to the performance of their subordinates and that the quality of that relationship depends on social exchange: "I as supervisor look after you, and you as subordinate perform your job well." Even when employees have little commitment to their employer as an organization, they may feel commitment to their supervisor. How is this exchange affected when decisions that had been made by the supervisor are now made by or even largely informed by an algorithm rather than a supervisor?

If my supervisor assigns me to work another weekend this month, something I very much do not want to do, I might do it without complaint if I think my supervisor has

Electronic copy available at: https://ssrn.com/abstract=3263878

otherwise been fair to me. I might even empathize with the bind my supervisor is in when having to fill the weekend shift. If not, I might well go complain to her and expect some better treatment in the future. When my work schedule is generated by software, on the other hand, I have no good will built up with that program, and I cannot empathize with it. Nor can I complain to it, and I may well feel that I will not catch a break in scheduling in the future. We know, for example, that people respond very differently to decisions that are made by algorithms than decisions made by people.[xlvi] If there is good news to give me, such as a bonus, it builds a relationship with my supervisor if she appears to have at least been involved in the decision, something that does not happen if that decision is generated by an algorithm.

Yet, there may be occasions where decisions are easier to accept when made by an algorithm than when made by a human, especially when those decisions have negative consequences for us. Uber riders, for example, respond negatively to surge pricing increases when they perceive that they are set by a human (trying to exploit them) as opposed to by an algorithm. Experimental evidence suggests that we are more willing to accept decisions from algorithms when we can see how they update to deal with mistakes.[xlvii]

Related to these issues is the engagement in decisions that individuals have that is otherwise lost with the shift to algorithms. Research increasingly shows that algorithms perform better than human judgment when used to predict repetitive outcomes, such as reading x-rays and predicting outcomes about employees or job candidates.[xlviii] But if algorithms take over hiring, and supervisors play no role in the process, will they be as committed to the new hires as if they had made the hiring decisions?

Discussion

In this section, we bring together our responses to the challenges to AI in HR management and refine them into more specific practical recommendations. We summarize our recommendations in Table 1 where the column "Operations" is placed at the end to reflect companies' need to respond to the reality modified by AI algorithms. A bird's-eye review of the table reveals three groups of recommendations across the stages of the AI Life Cycle that we label as causal reasoning, randomization and experiments, and employee contribution. Below we discuss them in more detail.

Electronic copy available at: https://ssrn.com/abstract=3263878

Table 1. Possible Responses to Challenges of AI's Introduction in Human Resource Management

Challenge	Response			
	Data Generation	**Machine Learning**	**Decision-Making**	**Operations**
Complexity of HR outcomes	Solicit employee contributions into outcomes' metrics and create consensus around them	Train algorithms for a few outcomes	Managers's discretion on the basis of the algorithm's predictions Run experiments whereby an algorithmic or human decision is randomly assigned to individual cases	Monitor the medium- and long-term validity of AI-based decisions Periodically review and retrain the algorithm
Small data	Integrate HR data with financial and operational data Use fine-grained real-time data Use vendors' data collected from larger populations	Use vendor-trained models Use causal models	Let managers act on algorithm's recommendations according to prespecified guidelines	
Accountability regarding fairness, ethical norms, and labor laws	Assess the consistency of human-made decisions used for training the algorithm Use	Create consensus around fairness criteria Weigh multiple fairness criteria Use causal model Ask data scientists to explain the model (identify the features that disproportionally affect its predictions)	Make random choices with probabilities predicted by the algorithm	Specify a code of ethics for AI-related initiatives Create an AI council with representatives of all stakeholders
Employee reactions	Collect data to improve processes first	Create employee consensus around the features used to train the algorithm	Maintain managers' responsibility for AI-based decisions Create an appeal process	Regularly solicit employee feedback Monitor employee engagement

Electronic copy available at: https://ssrn.com/abstract=3263878

Most machine learning based algorithms excel in pattern recognition by associations rather than *causation*. However, recognizing images, a common AI task, is not as difficult as recognizing good workers. As noted above, the scope of possible performance indicators is broad and hard to observe and measure precisely. Attempts to dig them out from digital traces of human behavior within and outside organizations run into severe issues of control, privacy, and ethics and still do not guarantee that anything worthwhile will be found. And even if tight associations between a set of observable worker characteristics and behaviors within a company is found, this set is unlikely to be fully usable on the pool of job candidates. Causal reasoning focuses our attention on the characteristics and behaviors that are relevant, reduces the costs of data management, and goes a long way towards meeting the requirements of fairness and explainability that are central to the future of AI algorithms.

The benefits of causal reasoning do come with costs. Causal models have lower predictive power in comparison with algorithmic associational models, although the evidence for this in the HR area is still unclear because we do not have evidence from HR tech vendors and proprietary algorithms on their validity. Causal models require not just data, but also expert knowledge of businesses as well as social scientists who should be hired together with data and computer scientists. By making causal assumptions explicit, algorithm designers make themselves vulnerable to criticism and organizational politics. We propose two remedies in this regard, methodological and organizational.

Methodologically, causal discovery is a rapidly developing toolkit that automates empirical testing of causal assumptions and thereby narrows down the set of plausible causal models for further consideration.[xlix]

Organizationally, the vulnerability of algorithm designers to criticism is a precondition for trust. We suggest that they channel criticism and politics through AI Councils that include widely respected representatives of all stakeholders, vigorously debate the assumptions, data, and ethical dimensions of AI-algorithms, and continuously solicit employees' contribution and feedback. Google moved in a related direction in March 2019 by announcing the Advanced Technology External Advisory Council to guide the company on how to ethically develop its AI technologies for business. The

Electronic copy available at: https://ssrn.com/abstract=3263878

Council lasted for only a week because of the controversy around one of its members. Some commentators of this debacle justly noted that "Google already has a tremendous resource in many of its own employees"[l] and, we would add, has to treat its employees as internal clients who deserve to consult when and how AI affect their work and careers.

The push towards causal modeling might face the resistance of organizational data scientists who rely on predominantly associational methods of algorithm training and do not follow emerging trends in computer science. However, we expect the trend towards causal algorithms to move rapidly from academic circles to the public arena as society confronts multiplying legal and ethical challenges of AI. Targeted learning[li] is one methodological approach that combines correlations-based pattern recognition with the subsequent targeted estimation of causal parameters and thus delivers a larger scope of benefits: more accurate predictions, generalizability, explainability, and fairness.

Randomization and experiments constitute our second principle that can help with algorithmic-based decisions. First, intentionally randomizing the inputs into an algorithm is akin to quasi-experimentation and can help to establish causality. Second, choosing an HR outcome, such as who gets the promotion, explicitly with a random component based on the probability predicted by an algorithm where we cannot predict outcomes with much accuracy acknowledges the inherently stochastic nature of HR outcomes and the unavoidable inaccuracy of algorithms. Employees may perceive such randomization—such as flipping a coin—to produce fairer outcomes under uncertainty. This is particularly important if some form of discrimination against legally-protected groups is intrenched in the organizational structures, processes, or culture.

Managers may decide to exercise their discretion over an algorithm's predictions and recommendations, in part to maintain control over outcomes and, in turn, over their subordinates. In this case, AI turns into augmented intelligence, which is the dominant modus operandi in data science today. To preserve the integrity of decision-making, this discretion should be a subject of some algorithm (e.g., in what step of the process are we using judgment) even of a back-of-the-envelope kind. Meehl's classical finding that simple statistical algorithms are better than human "clinical" judgment withstood the test of time and has been incorporated into the best HR decisions.[lii] Companies can solicit relevant employees' contributions into AI algorithms for a few performance criteria and

Electronic copy available at: https://ssrn.com/abstract=3263878

articulate ex ante how algorithms' quantitative outputs and human discretion ought to be combined in a final qualitative decision.[liii]

Throughout this section, we emphasized *employee contribution* as the third critical response to all the challenges of AI which is consistent with studies of algorithm aversion noted above. From the standpoint of organization theory, we see AI as an innovative organizational process whose introduction can be enabling rather than coercive[liv] if all stakeholders can participate in the AI Life Cycle. In addition to AI Councils proposed above, formal channels for appealing high-stake algorithmic decisions and for submitting feedback more generally have to be put in place. Vendors of cloud-based HR services are positioned the best to develop valid causal models for recruitment, from which all their clients can potentially benefit.

Whether clients are willing to let their data be aggregated for this opportunity to realize remains to be seen as does whether the vendors will take the high road in searching for the best and fairest decisions. Escalating our suggestion of AI Councils up "into the cloud" can be one way to let clients monitor vendors' handling of their data and decide what algorithms can be implemented. Debiasing current HR decisions is potentially the most valuable opportunity for AI in HR decision-making in the short- to medium-run which data scientists and companies should act on. The successful removal of human arbitrariness from HR decision-making should by itself lessen algorithmic aversion and make workers accustomed to AI-managed organizations.

In terms of guidance for practitioners, the arguments above suggest how to take on HR questions with data science rather than which tasks to take on. That advice suggests, though, that the most complicated and challenging HR task to address with data science techniques is likely to be hiring because so many fairness and legal issues are at play there. Given that, it might make more sense to start elsewhere, such as with natural language processing analyses of data such as that generated by open-ended questions in employee surveys and performance feedback conducted via apps. Finding patterns in responses would be extremely helpful, few organizations seem to try to find them now, and it is a straight-forward task for data science to provide these descriptive results.

The next set of tasks to take on might be those that involve machine learning algorithms but on topics that do not involve HR outcomes subject to legal and fairness

Electronic copy available at: https://ssrn.com/abstract=3263878

considerations. These include "advice" given to employees about training programs that make sense for them, new jobs for which they might apply, wellness and retirement planning advice, and so forth.

Before moving with machine learning into topics like recruitment and selection, dismissals, or promotion decisions, where legal and fairness questions are paramount, it makes sense to see how what we are doing now is working. That question is best answered with traditional statistical methods and hypothesis testing: do employee referrals actually provide the best candidates for us, do the personality tests we give predict good performers, do graduates from elite schools actually perform better in our jobs than other hires, etc. It also makes sense to investigate the confidence we have in any available job performance measures.

Conclusions

While the deployment of general-purpose AI is still a long shot in any domain of human activity, the speed of progress towards specialized AI systems in health care, automobile industry, social media, advertising and marketing has been considerable. Far less progress has been made in issues around the management of employees on the first step of the AI path, which is decisions guided by algorithms. We identify four reasons why: complexity of HR phenomena, data challenges from HR operations, fairness and legal constraints, and employee reactions to AI-management.

Causal reasoning is the first principle relevant to addressing these challenges across the stages of the AI Life Cycle as it helps greatly with issues of fairness and explainability. However, the benefits of causal reasoning come with costs. Employers must first accept the greater costs (based on the need for more data) and lower predictive power from their algorithms, and they must work to develop consensus about causal assumptions in advance of modeling. These challenges explain why the data science community is often skeptical about causally reasoning AI systems.

Randomization is a second principle that can help with algorithmic-based decisions, in part, recognizing that what constitutes "noise" in modeling is actually useful for improving algorithmic models. Our more surprising conclusion is that fairness and the explainability of HR decisions can be improved significantly by acknowledging when

Electronic copy available at: https://ssrn.com/abstract=3263878

our algorithms are not very valid and including an explicit element of randomness in the decision process.

We also recognize the limits of a top-down optimization approach to HR decisions because of the negative effects it is likely to have on employee behavior. Ensuring employee involvement in the process of building and using algorithms is a necessary for their success.

To what extent the changes we suggest require a restructuring of the HR function is an important question. Certainly, HR leaders need to understand and facilitate the Data Generation and Machine Learning stages of the AI Life Cycle, and new competencies may be needed to make that happen. The use of data analytics should help the HR function integrate more closely with other parts of the business, particularly finance and operations. There is a risk to HR leaders that if they do not engage the possibilities of AI, some other function in the business will take control of it for them.

Line managers will have to refresh their skill set as well. For them, AI should imply "augmented intelligence," an informed use of workforce analytics' insights in decision-making. The literature on evidence-based management proposes a Bayesian approach to systematically updating managerial beliefs with new information.[lv] We consider it a helpful departure point for AI-management as well.

The tension between the logic of efficiency and of appropriateness affects most organizational action, as March and Simon noted.[lvi] In the case of HR, the drive for efficiency and concerns about fairness do not always align. We hope that the conceptual and practical insights in this paper will move AI-management in HR forward on both counts, those of efficiency and appropriateness.

---

[i] IBM. Unplug from the Past: 19th Global C-Suite Study. IBM Institute for Business Value. 2018.
[ii] LinkedIn. *The Rise of HR Analytics*. 2108.
[iii] F.D. Schoorman, F. D. Escalation bias in performance appraisals: An unintended consequence of supervisor participation in hiring decisions. *Journal of Applied Psychology*, 1988. *73*(1), 58.

[iv] Peter Cappelli and Anna Tavis. "The performance management revolution." *Harvard Business Review* 94.10 (2016): 58-67.
[v] See Angelo DeNisi and C.E. Smith. Performance appraisal, performance management, and firm-level performance: A review, a proposed model, and new directions for future research. *Academy of Management Annals*, 2014. *8*(1), 127-179.
[vi] Peter Cappelli. There's no such thing as big data in HR. 2017. *Harvard Business Review*, *2*.

Electronic copy available at: https://ssrn.com/abstract=3263878

[vii] Enric Junque de Fortune, David Martens, and Foster Provost. Predictive Modeling with Big Data. Big Data. 1(4). December 13th 2004.

[viii] David Meyer. *Amazon Reportedly Killed an AI Recruitment System Because It Couldn't Stop the Tool from Discriminating Against Women.* Fortune. October 10th2018.

[ix] See Elizabeth Barends and Denise M. Rousseau. *Evidence-based management: How to use evidence to make better organizational decisions*. 2018. Kogan Page Publishers; Jeffrey Pfeffer and Robert Sutton. *Hard facts, dangerous half-truths, and total nonsense: Profiting From evidence-based management*. 2006. Harvard Business Press; and Denise ousseau(Ed.) *The Oxford handbook of evidence-based management*. 2012. Oxford University Press.

[x] See J. Pearl. *Causality*. 2009. Cambridge university press and J. Pearl and D. Mackenzie. *The book of why: the new science of cause and effect*. 2018. Basic Books.

[xi] See, e.g., Bo Cowgill. (2018). Bias and Productivity in Humans and Algorithms. Theory and Evidence from Résumé Screening. Working paper. J. Denrell, J., C. Fang, and C. Liu. Perspective—Chance explanations in the management sciences. *Organization Science*, 2014. *26*(3), 923-940. C. Liu and J. Denrell. *Performance Persistence Through the Lens of Chance Models: When Strong Effects of Regression to the Mean Lead to Non-Monotonic Performance Associations*. 2018. Working paper.

[xii] E.A. Lind and K. Van den Bos. When fairness works: Toward a general theory of uncertainty management. *Research in organizational behavior. 2002. 24*, 181-223.

[xiii] S. Barocas and A.D. Selbst. Big data's disparate impact.2016. *Calif. L. Rev.*, *104*, 671 and J. Kleinberg, J. Ludwig, S. Mullainathan, and A. Rambachan. Algorithmic fairness. 2018. In *AEA Papers and Proceedings* (Vol. 108, pp. 22-27).

[xiv] See Michael D. Giandrea and Shawn A. Sprague. Estimating the US Labor Share. 2017. Monthly Labor Review. February.

[xv] Human Resources Cycle: Comparison of Models. *The Oxford Review*. https://www.oxford-review.com/oxford-review-encyclopaedia-terms/human-resources-cycle/.

[xvi] Logistic regression refers to a supervised machine learning technique that is commonly used for predictive analysis. Logistic regression uses a logistic function to predict a binary outcome variable of interest.

[xvii] M. K. Lee, D. Kusbit, E. Metsky, and L. Dabbish.. Working with machines: The impact of algorithmic and data-driven management on human workers. 2015. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 1603-1612). ACM.

[xviii] S. Netessine and V. Yakubovich. The Darwinian workplace. 2012. *Harvard business Review*, *90*(5), 25.

[xix] Eric Rosenbaum. IBM artificial intelligence can predict with 95% accuracy which workers are about to quit their jobs. 2019. CNBC. https://www.cnbc.com/2019/04/03/ibm-ai-can-predict-with-95-percent-accuracy-which-employees-will-quit.html.

[xx] See, e.g., K.R. Murphy. Criterion issues in performance appraisal research: Behavioral accuracy versus classification accuracy.1991. *Organizational behavior and human decision processes*, *50*(1), 45-50 and L. Roberson, B.M. Galvin, and A.C. Charles. 13 When group identities matter: bias in performance appraisal. 2007. *The academy of management annals*, *1*(1), 617-650.

[xxi] Barends and Rousseau, op.cit.

[xxii] See C. Rosengren and M. Ottosson. Employee monitoring in a digital context. 2017. *Digital sociologies*, 181-194 for an overview of digital monitoring.

[xxiv] Lazlo Bock. *Work rules!: Insights from inside Google that will transform how you live and lead*. 2015. Twelve.

[xxv] P.L. Roth, P. Bobko, C.H. Van Iddekinge, and J.B. Thatcher, J. B. Social media in employee-selection-related decisions: A research agenda for uncharted territory. 2016. *Journal of Management*, *42*(1), 269-298.

[xxvi] Pearl and MacKensie, op cit.

[xxvii] See www.eugdpr.org.

[xxviii] Bradley Areheart and Jessica Roberts. GINA, Big Data, and the Future of Employee Privacy. *Yale Law Journal*. 128(3): 710-790.

[xxix] C. Dwork and A. Roth, A. *The Algorithmic Foundations of Differential Privacy*. 2014. Now Publishers. See page 5.

[xxx] see Birkeland et al, op cit for an example

[xxxi] See, e.g. https://www.jobtestprep.co.uk/lloydsbank.

[xxxii] Bo Cowgill "Bias and Productivity in Humans and Algorithms. Theory and Evidence from Résumé Screening." 2018. Working paper.

[xxxiii] Pearl, op cit.

[xxxiv] Ben Schneider. The people make the place. 1987. *Personnel psychology*, *40*(3), 437-453.

[xxxv] M. Spielkamp, M. Inspecting algorithms for bias. 2017. *Technology Review*, *120*(4), 96-98.

[xxxvi] For details on the relevant US legal requirements, see D.J.Walsh. *Employment law for human resource practice*. 2015, Part II. Nelson Education.

[xxxvii] J.R. Loftus, C. Russell, M.J. Kusner, and R. Silva, R. Causal reasoning for algorithmic fairness. 2018. *arXiv preprint arXiv:1805.05859*.

Electronic copy available at: https://ssrn.com/abstract=3263878

[xxxviii] For a critical review, see Ronald J. Karren and Larry Zacharias. Integrity tests: Critical Issues. 2007. *Human Resource Management Review* 17(2) 221-234.
[xxxix] Loftus op cit., p.7.
[xl] D. Malinsky and D. Danks. Causal discovery algorithms: A practical guide. 2018. *Philosophy Compass*, *13*(1), e12470.
[xli] Cowgill 2018, op cit.
[xlii] Lind and Van De Bos, op cit.
[xliii] See P. Stone, P. *The luck of the draw: The role of lotteries in decision making*. 2011. Oxford University Press.
[xliv] J. Bloomberg, J. Don't Trust Artificial Intelligence? Time to Open the AI Black Box. 2018. Forbes. *Retrieved October*, *23*, 2018.
[xlv] See https://www.darpa.mil/program/explainable-artificial-intelligence.
[xlvi] B.J. Dietvorst, J.P. Simmons, and C. Massey. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. 2016. *Management Science*, *64*(3), 1155-1170.
[xlvii] B.J. Dietvorst, J P. Simmons, and C. Massey. Algorithm aversion: People erroneously avoid algorithms after seeing them err. 2015 *Journal of Experimental Psychology: General*, *144*(1), 114.
[xlviii] For example, see Cowgill 2018 op cit.
[xlix] See Malinsky and Danks, op cit.
[l] https://www.technologyreview.com/s/613281/google-cancels-ateac-ai-ethics-council-what-next/
[li] See M.J. Van der Laan and S. Rose S. *Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies*. 2018. Springer.
[lii] Paul Meehl. *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence. 1954.* University of Minnesota Press.
[liii] For examples, see Daniel Kahneman. *Thinking Fast and Slow.* 2011 Farrar, Straus and Giroux: pp.230-231).
[liv] Paul Adler and B. Borys. Two types of bureaucracy: Enabling and coercive. 1996. *Administrative science quarterly*, 61-89.
[lv] Bardens and Rousseau, op cit.
[lvi] James March and Herbert A. Simon. *Organizations*. 1958. *NY: Wiley, New York*.

Electronic copy available at: https://ssrn.com/abstract=3263878